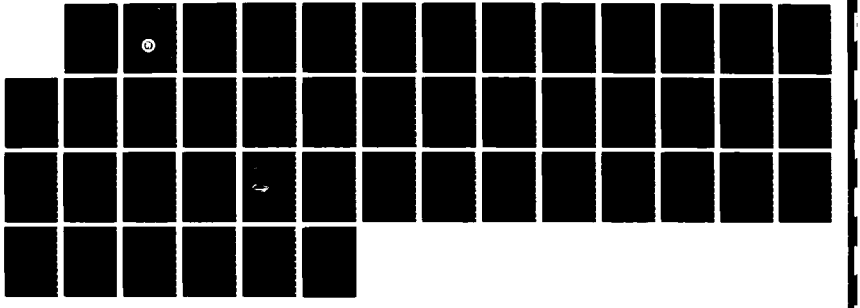
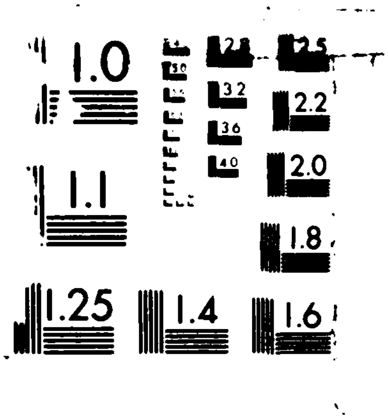
AD-A170 905 SOME INTEGRATED SQUARED ERROR PROCEDURES FOR MULTIVARIATE NORMAL DATA. (U) RENSSELAUER POLYTECHNIC INST TROY NY SCHOOL OF MANAGEMENT A S PAULSON ET AL. 1/1

UNCLASSIFIED 1986 ARO-18072.23-MA DAAG29-81-K-0110 F/G 12/1 NL





MIC

N

Rensselaer Polytechnic Institute
School of Management

Working Paper No. 37-86-P30

DISCLAIMER NOTICE

THIS DOCUMENT IS BEST QUALITY PRACTICABLE. THE COPY FURNISHED TO DTIC CONTAINED A SIGNIFICANT NUMBER OF PAGES WHICH DO NOT REPRODUCE LEGIBLY.

SOME INTEGRATED SQUARED ERROR PROCEDURES FOR MULTIVARIATE NORMAL DATA

by

A. S. Paulson, N. J. Delaney,
H. L. Hwang, and C. E. Lawrence

School of Management
Rensselaer Polytechnic Institute
Troy, NY 12180-3590
(518)266-6586

Note: This paper is not to be quoted without the permission of the author(s). For a list of Working Papers available from the School of Management, please contact Sheila Chao, School of Management.

Abstract

Two methods of estimation for the parameters of the multivariate normal distribution based on the sample characteristic function are given. These methods are shown to have an equivalent basis in terms of Parzen kernel-like density estimation. The estimators for the mean vector and covariance matrix are dependent on a user-specified parameter. Variation of the user-specified parameter produces a response surface in the parameter estimates and therefore allows for an informal sensitivity analysis of the data with respect to a tentative working model. The informal sensitivity analysis is intricately related to formal tests of fit of the model. The estimators of mean vector and covariance matrix have desirable robustness properties, are easy to compute and use, are relatively efficient at the multivariate normal, and are useful in identifying potential outliers and inconsistencies in some statistical assumptions. These methods are directly applicable to structured data such as multivariate experimental designs. Several illustrations are provided.

Some Integrated Squared Error Procedures for Multivariate Normal Data

by

A.S. Paulson*
Rensselaer Polytechnic Institute

N.J. Delaney
Northeastern University

E.L. Hwang
Oak Ridge National Laboratory

C.E. Lawrence
N.Y. State Department of Health

* Research sponsored in part by the U.S. Army Research Office under contract DAA G29-81-K-0110.

1. Introduction

This paper addresses the general problem of using sample characteristic functions to construct robust estimators of location and covariance parameters of linear models with a p-variate Gaussian error structure. We envision the linear models as tentative working models consisting of an error structure (additive Gaussianity) and a functional structure (e.g., a linear regression or experimental design model). Our procedures have also been used successfully on non-linear models but we do not address non-linear models here. The working model is regarded as a single entity. Of course, if it is certain that the functional structure is linear and that the error structure is additive, independent p-variate Gaussian then there would be little interest in estimators derived from the sample characteristic function because of the ready availability of maximum likelihood and least squares estimators. But it is precisely because of uncertainty concerning functional and error structure in almost all practical situations that alternatives to least squares and maximum likelihood become interesting. The working model should undergo a process of criticism (see Box, 1979, Daniel, 1978, and Paulson and Nicklin, 1983, for discussion) in order to determine the degree to which the available data and the tentative, working model are mutually consistent. A component of the process of criticism can be based on robust estimators: if robust estimators and, say, maximum likelihood estimators "agree" in the specification of a model in the sense that if estimators of unknown parameters are close, then the data and the model <u>may</u> be mutually consistent. Several difficulties associated with the last sentence need to be highlighted. First, just as is the case for tests of fit, a particular robust method may not be sensitive to certain types of departure from a working model and thus the use of the word <u>may</u>. Secondly, how can "agreement" between working model and the data be objectively assessed and how is

closeness of maximum likelihood and robust estimates to be objectively assessed? Any formal assessment of agreement is tantamount to a test of fit of the truth of the working model. This line of discussion strongly suggests an intimate connection of tests of fit with robust methods. If an objective assessment is not available, then robust methods should still be of interest since they generate an informal sensitivity analysis: different but sensible methods-apart from efficiency considerations regarding the methods - of extracting information from sample data relative to a given, tentative model should not lead to materially different summarizations or conclusions if model and data are mutually consistent. It is worth noting that in many practical settings the form of the error structure of a working model may be important only as an indication of repeatability or homogeneity. Gaussianity per se, for example, may be of little or no interest.

The estimation methods we develop for Gaussian working models are intimately related to density estimation, but are quite different in spirit from the work of Parzen (1962), and Watson and Leadbetter (1963). One of the main features of our work is the development of an objective function from which robust estimators of location and covariance are jointly determined. The estimation procedures developed herein compare favorably with those of Maronna (1976), Devlin et al. (1975), Campbell (1980), Gnanadesikan and Kettenring (1972) and Huber (1981, Chapter 8). The recent books by Huber (1981), Barnett and Lewis (1978), and Gnanadesikan (1977) provide excellent reviews and discussions of a major portion of the literature. Some of the univariate counterparts of this paper were considered by Paulson and Nicklin (1983).

2. Estimators Based on Minimization of an Objective Function

In fundamental papers Rosenblatt (1956), Parzen (1962), and Watson and Leadbetter (1963) considered the problem of estimating a density. While the approach taken in these papers is nonparameteric, we shall be concerned with density estimation in a parametric framework. We shall be concerned exclusively in this section with the problem of estimating the mean vector μ and the positive definite covariance matrix $D=(\sigma_{ij})$ of the p-variate normal (Gaussian) density

$$f(x) = f(x;\mu,D) = |2\pi D|^{-\frac{1}{2}}\exp(-\tfrac{1}{2}(x-\mu)^T D^{-1}(x-\mu)), \tag{2.1}$$

given a random sample $x_1, x_2, \ldots, x_n$, and given that the normal model is the tentative working model. Functional structure is incorporated subsequently. We shall say that the x_j have the distribution $N_p(\mu,D)$ for brevity. We use the expression sensitivity analysis because observational information processed under the aegis of the working model by different methods should not change the tentative results very much if the working model and the data are mutually consistent. If they are not mutually consistent, then substantial differences may result. It is not possible to give a completely unambiguous definition of sensitivity for every type of problem that may be encountered in practice. The judgment of the application area must always be incorporated into the process of assessing sensitivity.

Let

$$\psi_n(u) = n^{-1}\sum_{j=1}^{n}\exp(iu^T D^{-\frac{1}{2}}(x_j-\mu)) \tag{2.2}$$

be a sample characteristic function; note that when μ and D are specified

$$E(\psi_n(u)) = \psi(u) = \exp(-\tfrac{1}{2}u^T u),$$

where $i^2 = -1$, u is a $p \times 1$ vector of real numbers, and $D^{-\frac{1}{2}}$ represents the unique symmetric square root matrix of D. When μ and D are not known, $\psi_n(u)$ is not a statistic. However, the parameters μ and D of (2.1) may be estimated by making $\psi_n(u)$ and $\psi(u)$ match up in some sense. There are many ways in which this may be done but we present only that which we have found to be most theoretically as well as practically useful.

If we define

$$Q_n(\mu, D, m) = \int_{R_p} |\psi_n(u) - \psi(u)|^2 \exp(-m^2 u^T u)\, du \qquad (2.3)$$

$$= \int_{R_p} R(u)\, R^*(u)\, du = \int_{R_p} |R(u)|^2\, du,$$

where $*$ denotes complex conjugate, R_p represents p-dimensional Euclidean space, and the residual $\psi_n(u) - \psi(u)$ weighted by $\exp(-\tfrac{1}{2}m^2 u^T u)$ is

$$R(u) = \exp(-\tfrac{1}{2}(1+m^2)u^T u) - n^{-1} \sum_{j=1}^{n} \exp(iu^T D^{-\frac{1}{2}}(x_j - \mu) - \tfrac{1}{2}m^2 u^T u). \qquad (2.4)$$

The expression (2.4) represents a difference of characteristic functions whose inverse is

$$r(x) = g(x) - g_n(x)$$

where $g(x)$ is the spherical normal distribution with

$$g(x) = (2\pi(1+m^2))^{-\frac{1}{2}p} \exp(-\tfrac{1}{2}(1+m^2)^{-1} x^T x), \qquad (2.5)$$

and $g_n(x)$ is an estimator of $g(x)$ with

$$g_n(x) = n^{-1}(2\pi m^2)^{-\frac{1}{2}p} \sum_{j=1}^{n} \exp(-(2m^2)^{-1} z_j^T z_j), \tag{2.6}$$

$$z_j = x - D^{-\frac{1}{2}}(x_j - \mu).$$

The expression $g_n(x)$ is unbiased for $g(x)$ when the x_j are p-variate Gaussian. When the x_j are not Gaussian, $g_n(x)$ can differ substantially from $g(x)$ since each x_j has an influence in the estimate of the population density. Note that (2.6) is not the usual Parzen kernel density estimator. In fact, $g_n(x)$ has both parametric (being dependent on μ and D) and nonparametric features.

By the multidimensional version of Parsevel's theorem (Feller, 1966, Chapter 15)

$$Q_n(\mu, D, m) = \int_{R_p} |R(u)|^2 \, du = (2\pi)^p \int_{R_p} r^2(x) \, dx. \tag{2.7}$$

Estimators for μ and D, tentative on the correctness of the model (2.1), are given by minimizing $Q_n(\mu, D, m)$ over μ and D for a specified value of m, $0<m<\infty$. Explicit integration of (2.3) yields

$$Q_n(\mu, D, m) = \pi^{\frac{1}{2}p} \left\{ n^{-1} m^{-p} \sum_{j=1}^{n} \sum_{k=1}^{n} \exp\left[-\frac{1}{2m^2} Q_{jk} \right] \right.$$

$$\left. - 2(\tfrac{1}{2}+m^2)^{-\frac{1}{2}p} \sum_{j=1}^{n} \exp\left[-\frac{1}{2(1+2m^2)} Q_j \right] + n(1+m^2)^{-\frac{1}{2}p} \right\}, \tag{2.8}$$

$$Q_j = (x_j - \mu)^T D^{-1} (x_j - \mu), \tag{2.9}$$

$$Q_{jk} = (x_j - x_k)^T (2D)^{-1} (x_j - x_k). \tag{2.10}$$

From

$$\frac{\partial Q_n(\mu,D,m)}{\partial \mu} = 0, \quad \frac{\partial Q_n(\mu,D,m)}{\partial D} = 0,$$

we find that the estimators of μ and D, $\hat{\mu}(m)$, $\hat{D}(m)$, say, satisfy the implicit equations

$$\mu = \sum_{j=1}^{n} w_j(m)x_j, \tag{2.11}$$

$$D = k(m,n)BA^{-1}D , \tag{2.12}$$

where

$$A = \Sigma(x_j-\mu)(x_j-\mu)^T\exp[-(4m^2+2)^{-1}(x_j-\mu)^TD^{-1}(x_j-\mu)],$$

$$B = \Sigma\Sigma(x_j-x_k)(x_j-x_k)^T\exp[-(4m^2)^{-1}(x_j-x_k)^TD^{-1}(x_j-x_k)],$$

$$w_j(m) = a_j(m)/a.(m),$$

$$a_j(m) = \exp[-(4m^2+2)^{-1}(x_j-\mu)^TD^{-1}(x_j-\mu)],$$

$$b_{jk}(m) = \exp[-\tfrac{1}{2}(x_j-x_k)^T(2m^2D)^{-1}(x_j-x_k)],$$

$$a.(m) = \sum_{j=1}^{n} a_j(m),$$

$$b..(m) = \sum_{j\neq k}\sum b_{jk}(m),$$

$$k(m,n) = \frac{1}{2n}\left(1 + \frac{1}{2m^2}\right)^{\frac{1}{2}p+1} .$$

In (2.12) $\Sigma\Sigma$ runs over all j and k and Σ runs over all j, j, k = 1, 2,..., n. A convenient way of interpreting (2.12) is that, approximately, D is determined in such a way that the product of one estimate of the covariance matrix, B, say, with the inverse of another estimate of the covariance matrix, A, say, is apart from constants,

the identity matrix. Note that $\hat{\mu}(\infty) = \hat{\mu} = n^{-1}\sum_{j=1}^{n} x_j$ and $\hat{D}(\infty) = \hat{D} = n^{-1}\sum_{j=1}^{n}(x_j - \bar{x})(x_j - \bar{x})^T$, the usual method of moments and maximum likelihood estimates for μ and D respectively.

An estimator for D based on A in the absence of B may be developed in several ways. First, the expectation of A is

$$E(A) = n\left[\frac{1+2m^2}{2+2m^2}\right]^{\frac{1}{2}(p+2)} D$$

so that estimators for μ and D, say $\mu^*(m)$ and $D^*(m)$, are determined by the joint implicit equations

$$D = n^{-1}\left[\frac{2+2m^2}{1+2m^2}\right]^{\frac{1}{2}(p+2)} A \qquad (2.13)$$

and (2.11). Alternatively, another set of estimators for μ and D based on A in the absence of B, $\tilde{\mu}(m)$ and $\tilde{D}(m)$, say, can be determined from the implicit equations (2.11) and

$$D = \frac{2+2m^2}{1+2m^2}\,\frac{A}{a_1(m)}\,. \qquad (2.14)$$

Equation (2.14) arises from determining the constant k which makes

$$E(\Sigma\{k(x_j-\mu)(x_j-\mu)^T - D\}\, a_j(m)) = 0.$$

Maronna (1976) gives further details concerning these methods of construction of robust estimators for μ and D. Paulson (1986) has developed the estimators $\tilde{\mu}(m)$ and $\tilde{D}(m)$ from maximizing a generalized version of the log likelihood.

Estimators for D based on B in the absence of A can be derived in exactly the same way. Note that these estimators need not be paired with an estimator for μ. First, an estimator for D based on B alone, $D_1(m)$ say, may be developed from the implicit relationship

$$D = (2(n)(n-1))^{-1}\left[\frac{1+m^2}{m^2}\right]^{\frac{1}{2}(p+2)} B; \qquad (2.15)$$

alternatively, an estimator for D, $D^+(m)$ say, may be developed from the implicit relationship

$$D = \frac{1+m^2}{2m^2} \frac{B}{b..(m)} . \qquad (2.16)$$

We only explicitly use the estimators $\hat{\mu}(m)$,$\hat{D}(m)$, $\tilde{\mu}(m)$,$\tilde{D}(m)$ and the estimators $\tilde{\mu}(\lambda)$, $\tilde{D}(\lambda)$ of section 5 in this paper as these would be the ones that would be normally used in practice.

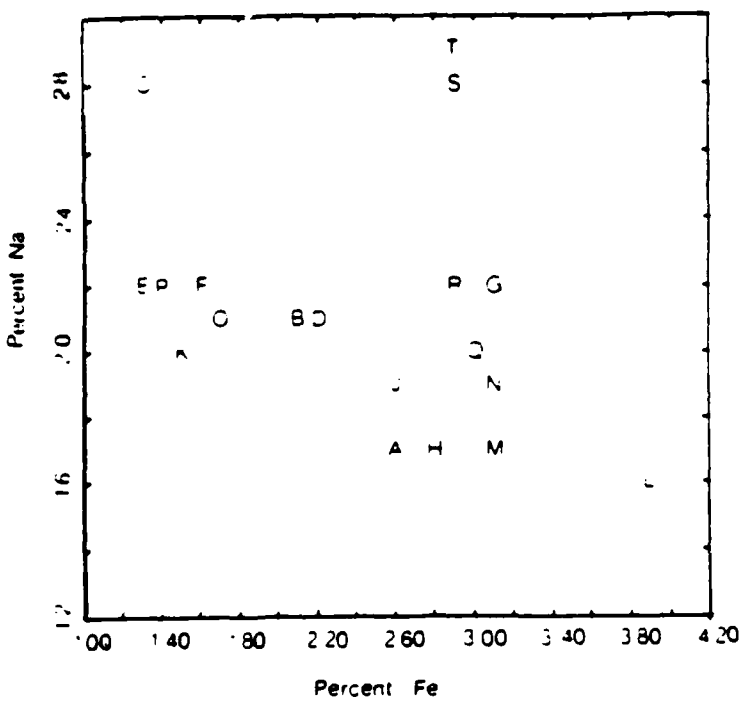
3. Numerical Computation of the Estimates Based on $Q_n(\mu,D,m)$

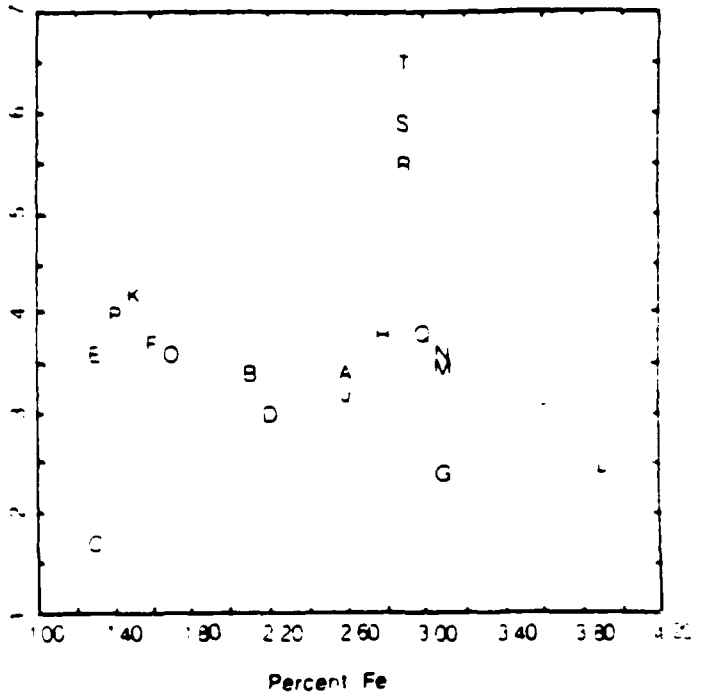
The expressions (2.11) and (2.12) are set out in accordance with a fixed point computational scheme. The left hand sides are designated as the updated estimates of μ and D; the right hand sides have current values of μ and D substituted wherever they appear. For example, an estimate of D would be computed via $D_{\ell+1} = k(m,n)B_\ell A_\ell^{-1}D_\ell$ where A_ℓ and B_ℓ are the ℓth iterates of A and B, each evaluated at the ℓth iterate of μ and D. Iteration proceeds until convergence is obtained. A Newton-Raphson scheme has also been used in place of the fixed point scheme but it is generally inferior to the fixed point scheme for $p \geq 3$. We have had good success in using $\hat{\mu} = n^{-1}\Sigma\, x_j$ and $\hat{D} = n^{-1}\Sigma\, (x_j-\hat{\mu})(x_j-\hat{\mu})^T$ as initial trial values of μ and D in (2.11) and (2.12). This strategy failed only, and infrequently at that, when the x_j in question had the character of two or more clusters of data. In this case it is clear that the tentative Gaussian model is inappropriate in any event. It is difficult to specify the numerical behavior of the estimation procedure because this behavior

depends to such an extent on the sample of x_j's in question, and the values of m, p, n, and the error tolerance on successive values of $Q_n(\mu,D,m)$. The greater p, the greater the number of iterations to convergence. For p = 1 and 2, m = 1, and convergence determined by successive values of $Q_n(\mu,D,m)$ being less than .001, convergence to a final solution was attained in fewer than 30 iterations for over ninety percent of a large number of trial problems of various sample sizes, $10 \leq n \leq 120$.

Example 3.1. The twenty triples (2.6, 1.7, 3.4), (2.1, 2.1, 3.4), (1.3, 2.8, 1.7), (2.2, 2.1, 3.0), (1.3, 2.2, 3.6), (1.6, 2.2, 3.7), (3.1, 2.2, 2.4), (2.8, 1.7, 3.8), (4.0, 1.4, 3.3), (2.6, 1.9, 3.2), (1.5, 2.0, 4.2), (3.9, 1.6, 2.5), (3.1, 1.7, 3.5), (3.1, 1.9, 3.6), (1.7, 2.1, 3.6), (1.4, 2.2, 4.0), (3.0, 2.0, 3.8), (2.9, 2.2, 5.5), (2.9, 2.8, 5.9), (2.9, 2.9, 6.5) represent the percentage of iron (Fe), sodium (Na), and potassium (K) obtained in a chemical analysis of twenty geological specimens. These are labeled sequentially as A, B,..., T. All two way scatterplots of this data given in Figure 1. Table 1 gives estimates of the components of $\hat{D}(m)$, $\bar{D}(m)$ $\tilde{D}(\lambda)$ (see Section 5 for defintion of $\tilde{D}(\lambda)$) for $m = +\infty$ (i.e., maximum likelihood), $m^2=2$, and $\lambda=2$.

The maximum likelihood estimators of the variances σ_{ii} and the correlations ρ_{ij} are drastically different from the estimators $\hat{\sigma}_{ii}(m)$, $\hat{\rho}_{ij}(m)$, $\bar{\sigma}_{ii}(m)$, $\bar{\rho}_{ij}(m)$, and $\tilde{\sigma}_{ii}(m)$, $\tilde{\rho}_{ij}(m)$ when $m^2 = 2$. Of particular interest is the value of the estimators of ρ_{23}. A case for the reasonableness of each of these estimators of ρ_{23} can be made depending on which observations one is willing to dismiss as being inappropriate to a Gaussian model. A close examination of Figure 1 will reinforce this point. The maximum likelihood estimator is not critical of the data in any sense while $\hat{D}(m)$, $\bar{D}(m)$, and $D^+(m)$ are effectively clustering the data in the way that each perceives will retain as much as possible the working Gaussian model. This clustering implies





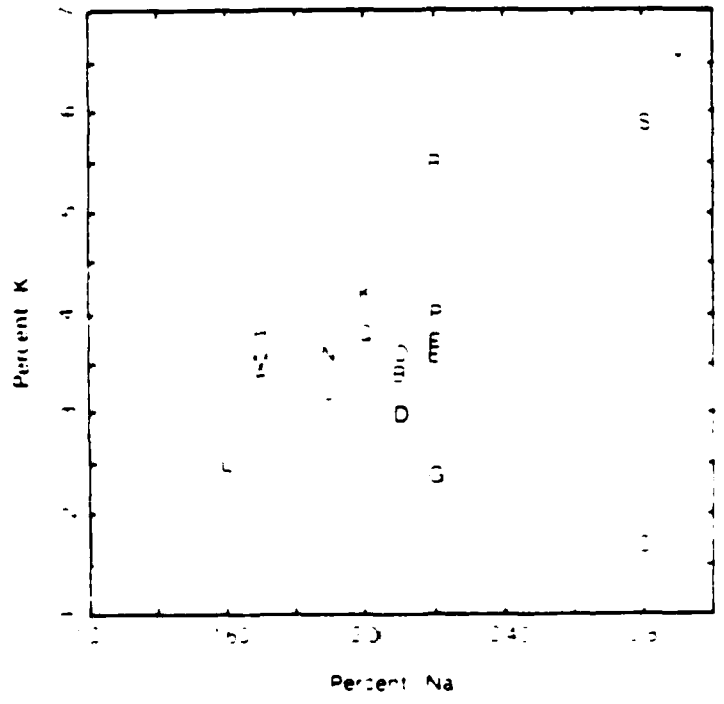


Figure 1. Two way scatterplots of percentages of Na, Fe, and K in 20 geological specimens of putatively the same origin.

that certain observations are being "filtered out" of the estimator of D by being heavily downweighted. This aspect of clustering receives more attention in the following sections.

This example highlights the fact that different estimators may focus on different aspects of the data under the aegis of some tentative working model. Accordingly we see that the parameter estimates are sensitive functions of the estimation procedure or of, also in this case, a parameter of the estimation procedure, m. Having seen such sensitivity, one must conclude that certain data is not consistent with the working model, e.g., perhaps they are outliers, or that the working model of independent, identically distributed Gaussianity is not appropriate for the setting at hand. Incidentially, the test of fit of Gaussianity of Paulson, Roohan, Hwang, and Fuller (1986) rejects this data as being Gaussian with p-value < .01.

Table 1

Parameter Estimates for Several Values of m^2

		Fe	Na	K			
Estimator	m^2	σ_{11}	σ_{22}	σ_{33}	ρ_{12}	ρ_{13}	ρ_{23}
$\hat{D}(m)$	2	1.19	0.166	0.633	-0.770	-0.133	-0.413
$\tilde{D}(m)$	2	0.776	0.070	0.229	-0.824	-0.508	0.137
$\tilde{\tilde{D}}(m)$	2	0.828	0.075	0.245	-0.828	-0.496	0.139
MLE	∞	0.646	0.149	1.23	-0.435	0.089	0.416

4. Statistical Properties of $\hat{\mu}(m)$, $\hat{D}(m)$

The simultaneously determined estimators $\hat{\mu}(m)$ and $\hat{D}(m)$ reduce, as $m \to \infty$, to the usual method of moments estimators $\hat{\mu} = n^{-1}\Sigma\, x_j$, $\hat{D} = n^{-1}\Sigma(x_j-\hat{\mu})(x_j-\hat{\mu})^T$. The estimators $\hat{\mu}(m)$ and $\hat{D}(m)$ are affine invariant. $\hat{D}(m)$ is positive definite with probability one for $m>0$ whenever $n>p$ although it may be algorithmically singular. The estimators $\hat{\mu}(m)$ and $\hat{D}(m)$ are consistent for μ and D for $m>0$ if $x_1, x_2, \ldots, x_n$ is a random sample from $N_p(\mu,D)$. The estimators $\hat{\mu}_1(m), \hat{\mu}_2(m), \ldots, \hat{\mu}_p(m)$ are jointly asymptotically normal for $m>0$ if $x_1, x_2, \ldots, x_n$ is a random sample from $N_p(\mu,D)$.

Explicit evaluation of $\partial Q_n(\mu,D,m)/\partial\mu$ shows that the estimator $\hat{\mu}(m)$ is an M-estimator for μ at the p-variate normal distribution. However, explicit evaluation of $\partial Q_n(\mu,D,m)/\partial D$ (and 2.12) shows that $\hat{D}(m)$ is dependent on pairwise differences x_j-x_k so that $\hat{D}(m)$ is not an M-estimator for D (Huber, 1981, pp. 43-44). Figures 2(a) - (f) provide influence function contours for the estimators $\hat{\mu}_1(1.5)$, $\hat{\sigma}_{11}(1.5)$, and $\hat{\sigma}_{12}(1.5)$ at the standard bivariate normal distribution for correlation $\rho=0$ and $\rho=.9$. In general, for $0<m<\infty$ the influence functions are bounded and redescendent to (matrix-valued, $p\times1$ or $p\times p$) zero as the Euclidean norm of the $p\times1$ vector argument of the influence function becomes unbounded. Thus both $\hat{\mu}(m)$ and $\hat{D}(m)$ are qualitatively robust estimators for μ and D for a fixed value of m. The finite sample multivariable sensitivity curves (see Barnett and Lewis, 1978, p. 137; Huber, 1981, pp. 15-16) $SC(x;\hat{\mu}(m),N_p(\mu,D))$ and $SC(x;\hat{D}(m),N_p(\mu,D))$ for $\hat{\mu}(m)$ and $\hat{D}(m)$ at $N_p(\mu,D)$ are also bounded and redescendent to (matrix-valued, $p\times1$ or $p\times p$) zero as the Euclidean norm of the $p\times1$ vector-valued argument x becomes unbounded for $n>p$ and $0<m<\infty$. Therefore, the contours of the sensitivity curves of the estimators $\hat{\mu}(m)$, $\hat{D}(m)$ are closed and bounded. This closedness property implies (1) that the process of estimation may be used as a clustering algorithm, and (2) that the process of

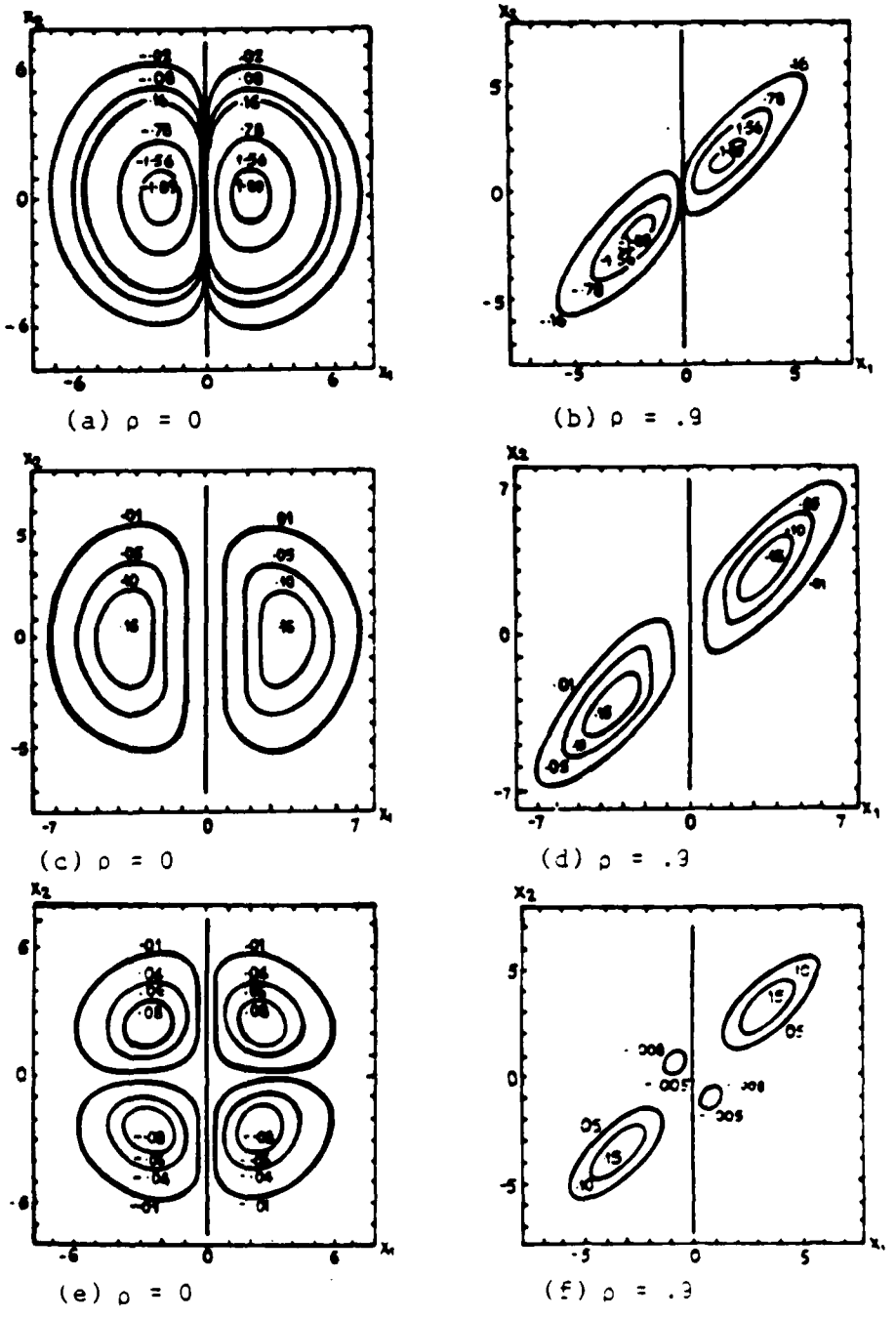


Figure 2. Influence functions for (a) $\hat{\mu}_1(m)$, $\rho = 0$, (b) $\hat{\mu}_1(m)$, $\rho = .9$, (c) $\hat{\sigma}_{11}(m)$, $\rho = 0$, (d) $\hat{\sigma}_{11}(m)$, $\rho = .9$, (e) $\hat{\sigma}_{12}(m)$, $\rho = 0$, (f) $\hat{\sigma}_{12}(m)$, $\rho = .9$ at the standard bivariate normal distribution with correlation ρ, $m^2 = 3.2$.

estimation may be used to evaluate the results of some clustering algorithms. The clustering capability associated with estimation of μ and D allows for identification of potential outliers in multivariate normal data.

The asymptotic efficiency of the j-th component of $\hat{\mu}(m)$, say $\hat{\mu}_j(m)$, relative to the sample mean is determined to be

$$\text{eff}(\hat{\mu}_j(m)) = \left[1 + \frac{c^2}{1+2c}\right]^{-\frac{1}{2}(p+2)}, \tag{4.1}$$

where $c = (1+2m^2)^{-1}$. The asymptotic efficiency of the j-th component of $\hat{D}(m)$, say $\hat{\sigma}_{jj}(m)$, relative to the sample variance is much more difficult to obtain but lengthy computations and extensive simulation trials suggest that

$$\text{eff}(\hat{\sigma}_{jj}(m)) \geq \left[1 + \frac{3c^2}{2+4c}\right]^{-1}\left[1 + \frac{c^2}{1+2c}\right]^{-\frac{1}{2}(p+2)}. \tag{4.2}$$

These computations also suggest that asymptotic efficiency of the off-diagonal components of $\hat{D}(m)$, say $\hat{\sigma}_{jk}(m)$, relative to the maximum likelihood estimator of σ_{jk} is bounded below by $\text{eff}(\hat{\sigma}_{jj}(m))$.

5. Modified Squared Error Estimation

The method of section 2 is not directly applicable to the structured data case, e.g., the cases of regression models and experimental layout models. Another, and more extensively applicable, sample characteristic function-based estimation procedure for Gaussian models is now developed. If $x_1, x_2, \ldots, x_n$ is putatively a random sample from $N_p(\mu, D)$, then

$$E(\exp(iu^T x_j)) = \phi(u) = \exp(iu^T\mu - \tfrac{1}{2}u^T Du). \tag{5.1}$$

Define the jth residual in u as

$$R_j(u) = \exp(iux_j) - \phi(u). \tag{5.2}$$

The sum of moduli squared of the residuals is given by

$$L(u) = \sum_{j=1}^{n} R_j(u)R_j^*(u) = \sum_{j=1}^{n} |R_j(u)|^2 . \tag{5.3}$$

The case of the sum of moduli squared of residuals closely parallels the usual sum of squares of residuals of least squares, the major difference being that L(u) depends on the nuisance parameter u as well as on the data. Information concerning μ and D in principle may be extracted from L by minimizing L over a sufficiently extensive fixed grid of nonzero values of u. Any such estimators for μ and D would then depend on the number of u-values as well as the location of these u-values and would not be affine invariant. This is too clumsy a state of affairs for practical applications and so another approach is necessary. If we multiply both sides of

$$\frac{\partial L}{\partial \mu} = 0, \qquad \frac{\partial L}{\partial D} = 0 \tag{5.4}$$

by $\exp(-\lambda u^T Du)$, $0<\lambda<\infty$, and integrate over R_p, estimators of μ and D will satisfy the implicit matrix valued equations

$$\int_{R_p} \frac{\partial L}{\partial \mu} \exp(-\lambda u^T Du)\, du = 0, \quad \int_{R_p} \frac{\partial L}{\partial D} \exp(-\lambda u^T Du)\, du = 0. \tag{5.5}$$

These estimators $\tilde{\mu}(\lambda)$ and $\tilde{D}(\lambda)$, say, of μ and D are dependent on the single

function $\exp(-\lambda u^T Du)$. Another benefit of multiplying (5.4) by $\exp(-\lambda u^T Du)$ is to make the resulting estimators affine invariant. After use of the matrix differentiation formulas of Dwyer (1967), the integrals in (5.5) may be explicitly evaluated and rearranged to provide the joint implicit estimating equations for μ and D, namely

$$\mu = \sum_{j=1}^{n} w_j(\lambda) x_j, \tag{5.6}$$

$$D = (1+2\lambda)^{-1} \sum_{j=1}^{n} w_j^*(\lambda)(x_j-\mu)(x_j-\mu)^T, \tag{5.7}$$

where

$$v_j(\lambda) = \exp(-\frac{1}{2} Q_j), \quad v.(\lambda) = \sum_{j=1}^{n} v_j(\lambda),$$

$$w_j(\lambda) = v_j(\lambda)/v.(\lambda),$$

and

$$w_j^*(\lambda) = \exp(-\tfrac{1}{2}Q_j) / \sum_{j=1}^{n} \left(\exp(-\tfrac{1}{2}Q_j) - \left[\frac{1+2\lambda}{2+2\lambda}\right]^{\frac{1}{2}(p+2)}\right),$$

$$Q_j = (x_j-\mu)^T((1+2\lambda)D)^{-1}(x_j-\mu) .$$

The joint estimators of μ and D determined by (5.6) and (5.7) may be computed by a fixed point algorithm with $\hat{\mu} = n^{-1}\Sigma x_j$ and $\hat{D} = n^{-1}\Sigma(x_j-\hat{\mu})(x_j-\hat{\mu})^T$ supplying the initial guesses of $\tilde{\mu}(\lambda)$, $\tilde{D}(\lambda)$. We have not found second order methods to be necessary in computing the estimators.

An application of the p-dimensional version of Parseval's theorem (Feller, 1966, Chapter 15) shows the equivalence of equations (5.5) to

$$2(2\pi)^p \sum_{j=1}^{n} \int_{R_p} \left\{ \frac{\partial f(x)}{\partial \mu} \overset{x}{*} f_\lambda(x) \right\} \{ f(x) \overset{x}{*} f_\lambda(x) - f_\lambda(x-x_j) \} \, dx = 0 \qquad (5.8)$$

and

$$2(2\pi)^p \sum_{j=1}^{n} \int_{R_p} \left\{ \frac{\partial f(x)}{\partial D} \overset{x}{*} f_\lambda(x) \right\} \{ f(x) \overset{x}{*} f_\lambda(x) - f_\lambda(x-x_j) \} \, dx = 0 \qquad (5.9)$$

respectively, where $h_1(x) \overset{x}{*} h_2(x)$ represents the convolution of $h_1(x)$ with $h_2(x)$. Note that

$$f(x) \overset{x}{*} f_\lambda(x) = |2\pi(1+\lambda)D|^{-\frac{1}{2}} \exp\left(-\tfrac{1}{2}(x-\mu)^T((1+\lambda)D)^{-1}(x-\mu) \right)$$

and that

$$f_\lambda(x-x_j) = |2\pi\lambda D|^{-\frac{1}{2}} \exp\left(-\tfrac{1}{2}(x-x_j)^T(\lambda D)^{-1}(x-x_j) \right) .$$

The remaining convolutions in (5.8) and (5.9) are given by

$$\frac{\partial f(x)}{\partial D} \overset{x}{*} f_\lambda(x) = \left\{ \frac{\partial}{\partial D} \left[f(x) \overset{x}{*} |2\pi\lambda K|^{-\frac{1}{2}} \exp\left(-\tfrac{1}{2}x^T(\lambda K)^{-1}x \right) \right] \right\} \Big|_{K=D} .$$

and

$$\frac{\partial f(x)}{\partial \mu} \overset{x}{*} f_\lambda(x) = \frac{\partial}{\partial \mu} \left\{ |2\pi(1+\lambda)D|^{-\frac{1}{2}} \exp\left(-\tfrac{1}{2}(x-\mu)^T((1+\lambda)D)^{-1}(x-\mu) \right) \right\} .$$

The equations (5.8) and (5.9) show that the estimators $\tilde{\mu}(\lambda)$ and $\tilde{D}(\lambda)$, may be derived, equivalently but less directly than $\hat{\mu}(m)$ and $\hat{D}(m)$, from considerations of parametric density estimation.

6. Properties of $\tilde{\mu}(\lambda)$, $\tilde{D}(\lambda)$

The estimators $\tilde{\mu}(\lambda)$ and $\tilde{D}(\lambda)$ are well-defined for $0<\lambda<\infty$. It is clear from (5.6) that $\tilde{\mu}(\lambda) \to n^{-1}\sum x_j$ as $\lambda\to\infty$. Explicit evaluation of (5.5) show that $\tilde{\mu}(\lambda)$ and $\tilde{D}(\lambda)$ are M-estimators for μ and D (Huber, 1981, pp. 43-44). The estimators are affine invariant. Slight modification of the arguments of Bryant and Paulson (1979) show that $\tilde{\mu}(\lambda)$ and $\tilde{D}(\lambda)$ are consistent for μ and D when the x_j constitute a random sample from $N_p(\mu,D)$. The estimators $\tilde{\mu}_1(\lambda)$, $\tilde{\mu}_2(\lambda),\ldots,$ $\tilde{\mu}_p(\lambda)$, $\tilde{\sigma}_{11}(\lambda)$, $\tilde{\sigma}_{12}(\lambda),\ldots,$ $\tilde{\sigma}_{1p}(\lambda)$, $\tilde{\sigma}_{22}(\lambda),\ldots,$ $\tilde{\sigma}_{2p}(\lambda),\ldots,$ $\tilde{\sigma}_{pp}(\lambda)$ are jointly asymptotically normal and the $\tilde{\mu}_j(\lambda)$ are asymptotically mutually independent of the $\tilde{\sigma}_{jk}(\lambda)$ if the x_j constitute a random sample from $N_p(\mu,D)$.

If we define the asymptotic efficiency of $\tilde{\mu}(\lambda)$ relative to $\hat{\mu}$ as the ratio of the determinant of their covariance matrices, it can be shown that

$$\text{eff}(\tilde{\mu}(\lambda)) = \left[\frac{3+8\lambda+4\lambda^2}{4+8\lambda+4\lambda^2}\right]^{\frac{1}{2}p(p+2)} . \tag{6.1}$$

Similarly, the asymptotic efficiency of the estimator $\tilde{\sigma}_{kk}(\lambda)$ of the kth diagonal element of D, σ_{kk}, relative to its maximum likelihood estimator can be shown to be

$$\text{eff}(\tilde{\sigma}_{kk}(\lambda) = \frac{9}{2}\,\frac{\left[\frac{1}{1+2\lambda}\right]^2\left[\frac{1+2\lambda}{2+2\lambda}\right]^{p+4}}{\left[\frac{1+2\lambda}{3+2\lambda}\right]^{\frac{1}{2}p}\frac{6+8\lambda+4\lambda^2}{(3+2\lambda)^2} - \left[\frac{1+2\lambda}{2+2\lambda}\right]^{p+2}} \tag{6.2}$$

$$\sim \frac{9}{p+8} \text{ as } \lambda \to \infty.$$

Selected values of these efficiences are given in Table 2. We have not explicitly calculated the efficiencies of $\hat{\sigma}_{jk}(\lambda)$ relative to its corresponding maximum likelihood estimator.

Table 2

Asymptotic efficiencies of $\hat{\mu}_j(\lambda)$ (first tabular entry) and $\hat{\sigma}_{jj}(\lambda)$ (second tabular entry) for selected values of λ and p

	λ				
	.5	1	2	4	∞
p					
	.84	.91	.96	.99	1
1					
	.78	.87	.94	.98	1
	.79	.88	.95	.98	1
2					
	.68	.77	.84	.88	.90
	.74	.85	.93	.98	1
3					
	.59	.69	.76	.80	.82
	.70	.82	.92	.97	1
4					
	.52	.62	.69	.73	.75
	.55	.72	.87	.95	1
8					
	.46	.56	.63	.67	.69

In the case p=1, eff($\hat{\sigma}_{kk}(\lambda)$) tends to unity with increasing λ but for p>1, eff($\hat{\sigma}_{kk}(\lambda)$ is bounded away from unity. The efficiency eff(($\hat{\sigma}_{kk}(\lambda)$) is monotone increasing with λ. Thus, the higher the dimensionality, the larger the value of λ one should use if efficiency is a major consideration in the choice of λ for estimation. We shall address this issue in detail in section 7.

The nature of $\tilde{D}(\lambda)$ as $\lambda\to\infty$ is determined by an application of L'Hospital's rule. Some elementary manipulations yield the implicit relationship

$$D = \frac{\sum_{j=1}^{n}(x_j-\dot{\mu})(x_j-\dot{\mu})^T}{\frac{1}{2}n(p+2)-\frac{1}{2}\sum_{j=1}^{n}(x_j-\dot{\mu})D^{-1}(x_j-\dot{\mu})} \tag{6.3}$$

that the estimator must satisfy asymptotically in λ. When $p=1$ (6.3) may be rearranged to get the usual moment estimator for D, but for $p\geq 2$ (6.3) cannot be so rearranged. The estimator of D defined by (6.3) does not seem to be otherwise interesting.

The influence functions for $\tilde{\mu}(\lambda)$ and $\tilde{D}(\lambda)$ at the p-variate normal are similar to those of $\dot{\mu}(m)$ and $\dot{D}(\lambda)$. The major difference is that the influence function $IF(x;\tilde{D}(\lambda),N_p(\mu,D))$, $0<\lambda<\infty$, while being bounded and redescendent, is not redescendent to zero, but rather to a positive definite matrix constant, as the Euclidean norm of x becomes arbitrarily large.

7. Choice of λ

There are two possible uses for the procedures we propose here. (We shall restrict our attention to the λ-procedure although there is a parallel development for the m-procedures.) The first is to specify a single value of λ, possibly based on efficiency considerations, and use it as a robust procedure. The choices $\lambda=1$ or $\lambda=2$ provide high efficiencies and good robustness properties. The second use is the one for which the procedure was developed and which has proved most useful in practical applications. We use the procedure to generate a sensitivity analysis. In a practical

exploratory setting we first compute the maximum likelihood estimators and use these for starting values in the iterative algorithm. Next we take a value of λ, 4 or 2, and examine the behavior of the estimates. Finally, we would take λ=1 or ½ and examine the behavior of the estimates. It is not possible to completely specify the values of λ because the behavior of the estimates depends on the nature and the extent of the data. The response surface of the parameter values and the final weights as a function of λ are of primary interest. In the process we determine estimates and final weights $\nabla_j(\lambda) = \exp(-\frac{1}{2}(1+2\lambda)^{-1}(x_j-\tilde{\mu}(\lambda))^T\tilde{D}(\lambda)^{-1}(x_j-\tilde{\mu}(\lambda)))$ associated with each observation. If the estimates are sensitive to this variation in λ, then there may be problems associated with either the data or with the Gaussian error model or both. It will not always be possible to determine the source of the problem. The particular observation(s) which is (are) the potential cause of the sensitivity are identified by low values of $\nabla_j(\lambda)$ vis-a-vis the whole set of these weights. This discussion will be subsequently clarified with an example.

The derivation which led to the estimators given in (5.6)-(5.7) did not require that λ in (5.6) be the same as in (5.7). We could, for example, use λ=1 for $\tilde{\mu}$ and λ=2 for $\tilde{D}$. Furthermore, we need not have restricted ourselves to a scalar value of λ in order to arrive at (5.6) and (5.7). At the expense of greater algorithmic complexity we could have chosen values λ_{ij} corresponding to each σ_{ij} in the covariance matrix D. Because D is symmetric, we take $\lambda_{ij} = \lambda_{ji}$. Let the p×p matrix $L = (1+2\lambda_{ij})$ and the p×p matrix $M = (2+2\lambda_{ij})$ and let $L \times D = ((1+2\lambda_{ij})\sigma_{ij})$ denote the Hadamard product of L with D. Then by arguments similar to those employed to arrive at (5.6) and (5.7) it may be shown that the more general estimators for μ and D satisfy the implicit relations

$$\mu = \frac{\sum_{j=1}^{n} x_j \exp(-\tfrac{1}{2}(x_j-\mu)^T(LxD)^{-1}(x_j-\mu))}{\sum_{j=1}^{n} \exp(-\tfrac{1}{2}(x_j-\mu)^T(LxD)^{-1}(x_j-\mu))}, \qquad (7.1)$$

and

$$LxD = \left\{ \sum_{j=1}^{n} \exp(-\tfrac{1}{2}(x_j-\mu)^T(LxD)^{-1}(x_j-\mu)) \right\}^{-1} \left\{ n \frac{|LxD|^{\frac{1}{2}}}{|MxD|^{\frac{1}{2}}} (LxD)(MxD)^{-1}(LxD) \right.$$

$$\left. + \sum_{j=1}^{n} (x_j-\mu)(x_j-\mu)^T \exp(-\tfrac{1}{2}(x_j-\mu)^T(LxD)^{-1}(x_j-\mu)) \right\}. \qquad (7.2)$$

The identity $MxD = \left[\frac{2+2\lambda_{ij}}{1+2\lambda_{ij}}\right] x\ (LxD)$ is useful in computing (7.2). The estimators of σ_{ij} are computed in a component-wise fashion from the final iteration of (7.2). These estimators would be of interest when it is desired to treat different components of the x_j differently.

8. Examples

Example 8.1. The basic data for this example are taken from Anderson (1958). The first 25 points consist of the first two (of four) components of this data with five additional (outlying) observations appended. We have chosen λ=4, 2, 1, ½ for this illustration. Table 3 summarizes the weights $v_j(\lambda)=\exp(-\frac{1}{2}(x_j-\tilde{\mu})^T((1+2\lambda)\tilde{D})^{-1}(x_j-\tilde{\mu}))$ associated with each point on the assumption that the data follow a single multivariate Gaussian distribution. Table 4 provides the estimates of the means and covariances as well as the maximum likelihood estimators. With $\lambda = +\infty$, all weights are the same. As λ decreases from $+\infty$, the weights become differentiated.

The 5 outlying observations are rendered distinctive by their diminishing weights $\nabla_j(\lambda)$ as λ decreases. This indicates that these observations are not consistent with the remainder of the observations and the assumption of a single Gaussian distribution.

Table 3

Sensitivity of Observational Weights
$\nabla_j(\lambda)(\times 100)$ to Variation in λ

			λ			
Point	x_1	x_2	4	2	1	.5
1	179	145	36	39	45	50
2	201	152	33	36	23	16
3	185	149	37	41	47	55
4	188	149	37	40	44	49
5	171	142	34	35	39	39
6	192	152	37	39	44	48
7	190	149	37	39	41	42
8	189	152	37	40	47	53
9	197	159	35	36	39	36
10	187	151	37	41	47	55
11	186	148	37	40	45	50
12	174	147	35	37	38	36
13	185	152	37	41	46	50
14	195	157	36	38	42	43
15	187	158	35	36	30	20
16	161	130	27	23	17	8
17	183	158	34	34	22	10
18	173	148	35	36	33	27
19	182	146	37	40	45	50
20	165	137	31	30	30	25
21	185	152	37	41	46	50
22	178	147	36	39	45	49
23	176	143	36	38	42	45
24	200	158	34	35	37	36
25	187	150	37	41	47	54
26	200	130	18	6	0	0
27	200	135	23	11	0	0
28	165	160	24	13	1	0
29	195	170	28	22	6	1
30	220	170	23	18	13	6

Table 4

Sensitivity of Parameter Estimates to Variation in λ

	λ = 4	2	1	.5
μ_1	185.5	185.2	184.9	184.9
μ_2	150.1	150.3	149.9	149.7
σ_{11}	148.1	148.8	155.4	136.6
σ_{22}	80.2	74.1	65.7	52.3
ρ_{12}	.52	.67	.85	.87

From (5.8) or (5.9) we find that estimation of μ and D is intimately related to the estimation of

$$f(x) \overset{*}{} f_\lambda(x) = |2\pi(1+\lambda)D|^{-\frac{1}{2}} \exp\left[- \frac{1}{2(1+\lambda)} (x_j-\mu)^T D^{-1}(x_j-\mu)\right]$$

by

$$\hat{f}_\lambda(x) = n^{-1} \sum_{j=1}^{n} |2\pi\lambda\tilde{D}(\lambda)|^{-\frac{1}{2}} \exp\left[- \frac{1}{2\lambda} (x-x_j)^T \tilde{D}^{-1}(\lambda)(x-x_j)\right] .$$

The density $f(x) \overset{*}{} f_\lambda(x)$ may be regarded as a tentatively posited (prior) distribution and $\hat{f}_\lambda(x$ as its estimate, given the x_j's. If the x_j are from $N_p(\mu,D)$, then $\hat{f}_\lambda(x) \cong \hat{f}_{\lambda'}(x)$, for all reasonable pairs λ, λ' ($\lambda \neq \lambda'$, $0 < \lambda, \lambda' < \infty$. If we took $\lambda = 1$ and $\lambda' = .001$, $\hat{f}_\lambda(x)$ and $\hat{f}_{\lambda'}(x)$ would not be similar unless n were large since then $\hat{f}_{\lambda'}(x)$ is approaching a sum of Dirac delta functions. If the x_j are not from $N_p(\mu,D)$ but from $h(x)$, say, then $\hat{f}_\lambda(x)$ will be an estimator for $h(x) \overset{*}{} f_\lambda(x)$ and $\hat{f}_\lambda(x)$ can be quite different from $f(x) \overset{*}{} f_\lambda(x)$. We thus suggest that the estimators $\hat{\mu}(\lambda)$ and $\tilde{D}(\lambda)$ and $\hat{\mu}(m)$ and $\hat{D}(m)$ are robust primarily because of their intimate relationship to goodness of fit through a parametric density estimation. Equations (5.8) and (5.9) imply the data is being adaptively screened so

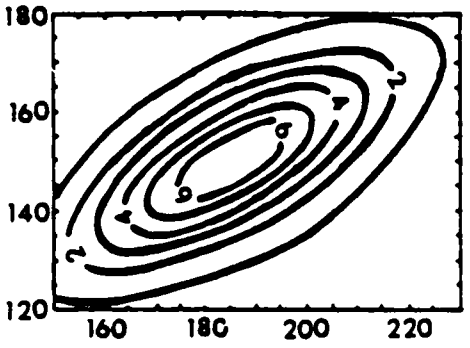
as to retain the character of Gaussainty as much as possible and that the parameter estimation is really subordinate to density estimation or error reconstruction.

Figures 3a, 3b, 3c, depicts what the estimation procedure perceives as λ decreases. For large λ the density estimate $\hat{f}_\lambda(x)$ is approximately uniform. At $\lambda=2$ the density estimate contours are smooth except for some slight distortion in the area of (195,130). At $\lambda=1$ the probability surface is distorted in the vicinity of the contamination but the distortion is not yet pronounced. Compare the estimates of the parameters and the observation weights $\nabla_j(\lambda)$. At $\lambda=\frac{1}{2}$ the distortion has become dramatic and indeed separate "hills" for the outlying points have formed. Again compare the estimates and $\nabla_j(\lambda)$ for $\lambda=\frac{1}{4}$. Since the density estimate perceives the outlying observations as not consistent with the remainder of the data and the single Gaussian assumption, the reason for the down-weighting of the outlying observations has become clear. If we let $\lambda \to 0+$, the density estimator becomes an average of a set of Dirac delta functions located at each point.

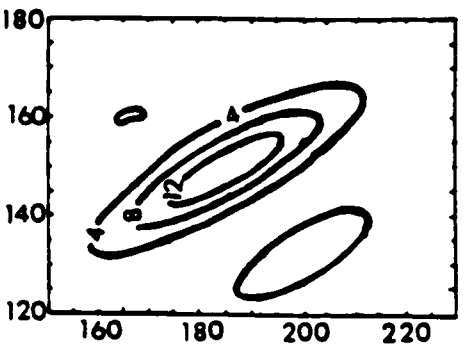
At $\lambda=\frac{1}{4}$, the procedure has effectively clustered the data with the outlying observations excluded from the main cluster. Accordingly, as λ is varied in this example, a dramatic change in the estimators implies the existence of clusters of observations different from the main cluster and not consistent with the prior assumption of independent, identically distributed Gaussianity. The reconstruction of the error density becomes increasingly critical with decreasing λ. This example reinforces the connection of the robust procedure with goodness of fit.

Example 8.2. If (y_j, z_j), $j = 1,2,\ldots, n$ represents a random sample from $N_2(\mu,D)$, $\mu = (\mu_1,\nu)^T$, then the regression of y on z is given by

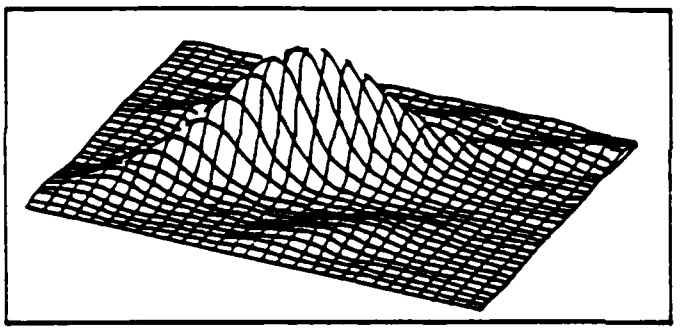
$$E(y|z) = \mu_1 + \sigma_{12}\sigma_{22}^{-1}(z-\nu) = \beta_0 + \beta_1 z,$$



(a) $\lambda = 2$



(b) $\lambda = 1$



(c) $\lambda = \frac{1}{2}$

Figure 3. Contour plots of $\hat{g}_\lambda(x)$ for (a) $\lambda = 2$ and (b) $\lambda = 1$. (c) High resolution density plot for $\hat{g}_\lambda(x)$, $\lambda = \frac{1}{2}$.

say, and the β's may be computed from the estimates of μ and D. The data for this example are taken from Andrews and Pregibon (1978) who were concerned with regression models. The data are presented in Table 5. The least squares or maximum likelihood estimates are also presented in Table 6. We shall use the L matrix version of the modified integrated squared error procedure as discussed in section 7.

The results concerning the variation in the $V_j(L)$ as a function of λ_{ij} or L are given in Table 5. The results concerning parameter estimates as a function of the λ_{ij} are given in Table 6.

Table 5

Sensitivity $\hat{V}_j(L)$(x100) to Variation in λ_{11}, λ_{12}, λ_{22}

			$(\lambda_{11}, \lambda_{12}, \lambda_{22})$		
	y	x	(2,4,8)	(1,2,4)	(.5,1,2)
1	95	15	100	99	97
2	71	26	72	48	20
3	83	10	85	76	65
4	91	9	96	93	88
5	106	15	88	82	74
6	87	20	96	89	72
7	93	18	99	96	87
8	100	11	98	97	97
9	104	8	95	91	84
10	94	20	96	91	78
11	113	7	81	70	54
12	96	9	99	97	95
13	83	10	85	76	65
14	84	11	88	81	71
15	102	11	97	95	93
16	100	10	98	97	96
17	105	12	92	89	84
18	57	42	40	11	0
19	121	17	50	33	19
20	86	11	92	86	79
21	100	10	98	97	96

Table 6

Sensitivity of Estimates to Variation in λ_{11}, λ_{12}, λ_{22}

(∞,∞,∞)	(∞,∞,∞)	(2,4,8)	(1,2,4)	(.5,1,2)
μ_1	14.4	13.4	12.8	12.3
ν	93.9	94.6	95.3	95.9
σ_{11}	60.1	45.6	32.7	23.6
σ_{12}	-81.2	-49.2	-29.2	-14.8
σ_{12}	190.2	168.3	143.9	125.1
ρ_{12}	-.76	-.56	-.43	-.27

The parameter estimates are sensitive functions of L. The points which are most influential or potentially inconsistent vis-a-vis the linear model with a Gaussian error

structure are determined from observational weights $\nabla_j(L)$, i.e., those with low values of $\nabla_j(L)$. Three points, 2, 18, 19, are especially singled out. Point 18 represents an extreme point in the z-space and is most influential on the estimate of the slope $\beta_1 = \sigma_{12}/\sigma_{22}$ of the regression line. Point 2 has the second-most extreme value of z. Point 19 produces an extreme residual. Analogous results obtain if all $\lambda_{ij} = \lambda$ and λ is varied in order to determine the response of the parameter estimates and the observational weights to variation in λ. Practical experience shows that in many, but not all, cases, taking just a single value of λ (or set of values λ_{ij}) provides sufficient information concerning the response of the parameter estimates and weights $\nabla_j(\lambda)$ (or $\nabla_j(L)$) to variation in λ (or L) in the sense that if λ (the λ_{ij}) were further decreased, the same trend in response will be continued. In these cases a robust analysis will lead to the same conclusions as the sensitivity analysis. In some cases, a change in trends will be observed as λ decreases so that a single robust analysis vis-a-vis maximum likelihood may not give the same indications as a sensitivity analysis.

Example 8.3. The three dimensional data for this illustration are taken from Gnanadesikan (1977, pp. 50-52). The 61 triads of his example 7 were obtained by adding a spherical Gaussian noise component to each of the coordinates on the surface of a specified paraboloid. The two-dimensional scatter plots of these data are not suggestive of the data in three dimensions lying near a curved surface. We determine the response of the parameter estimates to changes in λ. The mean vector estimate at $\lambda=0$ is (-3.54, 4.72, 26.94) while at $\lambda=\frac{1}{4}$ it is (-3.53, 4.70, 26.95). The variance estimates at $\lambda=0$ are (3.81, 2.33, 2.94) while at $\lambda=\frac{1}{4}$ they are (4.43, 2.76, 3.79); the correlation estimates at $\lambda=0$ are (-.51, -.47, .20) while at $\lambda=\frac{1}{4}$ they are (-.56, -.50, .27). The estimates of the mean are remarkably stable but the estimated variances and absolute values of the correlation increase with a

decrease in λ. These characteristics imply that if the Gaussian distribution model is not appropriate, the best place to look for difficulties is at the centroid. This is confirmed by the distribution of the weights $v_j(\lambda)$, especially for $\lambda=4$. For example, for $\lambda=4$, the largest three weights $v_j(\lambda)$, .89, .84, .82, indicate that there are no observations near the centroid. This deficiency of observations near the centroid is also determinable from $\lambda=0$ results, but it is highlighted at the smaller values of λ. The practice of examining the $v_j(\lambda)$ is similar to that of examining quantile-quantile plots (Gnanadesikan, 1977, pp. 50-52) since $v_j(\lambda)$ is based on the quadratic form $(x_j-\mu(\lambda))\tilde{D}^{-1}(\lambda)(x_j-\mu(\lambda))$. The conclusions drawn for this example are the same.

This informal observation of sensitivity of variance and covariance estimates to changes in λ can be made formal so as to provide probability statements concerning appropriateness of the p-variate normal model by defining a test of fit based on the nonnegative statistic

$$S(\lambda) = \mathrm{tr}(\dot{D}^{-1}(\lambda)\dot{D}) + \mathrm{tr}(D(\lambda)\dot{D}^{-1}))-2p.$$

for $\lambda = 1$, say. This test statistic will be near zero if Gaussianity is appropriate; it will be large if Gaussianity is not appropriate. Paulson and Swope (1986) have used a statistic similar to $S(\lambda)$ to test for p-variate normality. This example again shows the interrelationship of some aspects of robustness and sensitivity analysis with tests of fit of models.

9. Multivariate Two-Way Cross Classification

The system of equations (5.5) are readily extended to include multivariate regression and design situations. We indicate how this may be done for the case of a

two-way cross classified design. The arguments are similar for other designs and regression problems.

The two-way cross classified model may be written

$$E(x_{jkl}) = \mu + \alpha_j + \beta_k \ , \tag{9.1}$$

$j = 1,2,\ldots,a$, $k = 1,2,\ldots,b$, $l = 1,2,\ldots,n_{jk} \geq 1$, where the x_{jkl} are assumed to be p-dimensional Gaussian with covariance matrix D. The quantities μ, α_j, β_k are px1 location vectors. Define

$$\phi_{jk}(u) = \exp(iu^T(\mu+\alpha_j+\beta_k) - \tfrac{1}{2}u^TDu)) \tag{9.2}$$

and

$$L(u) = \sum_{j=1}^{a} \sum_{k=1}^{b} \sum_{l=1}^{n_{jk}} |\phi_{jk}(u) - \exp(iu^Tx_{jkl})|^2 \ . \tag{9.3}$$

Multiply both sides of each of the following

$$\frac{\partial L}{\partial \mu} = 0, \ \frac{\partial L}{\partial \alpha_j} = 0, \ \frac{\partial L}{\partial \beta_k} = 0, \ \frac{\partial L}{\partial D} = 0 \ , \tag{9.4}$$

by $\exp(-\lambda u^TDu)$ and integrate over R_p to get

$$\int_{R_p} \frac{\partial L}{\partial \theta} \exp(-\lambda u^TDu)\, du = 0 \tag{9.5}$$

for $\theta = \mu, \alpha_j, \beta_k, D$, $j = 1,2,\ldots, a$, $k = 1,2,\ldots, b$. Explicit evaluation of (9.5) leads to the system of implicit equations

$$\sum_j \sum_k \sum_l (x_{jkl}-\mu-\alpha_j-\beta_k)\, v_{jkl}(\lambda) = 0,$$

$$\sum_k \sum_l (x_{jkl}-\mu-\alpha_j-\beta_k)\, v_{jkl}(\lambda) = 0, \quad j = 1,2,\ldots, a$$

$$\sum_j \sum_l (x_{jkl}-\mu-\alpha_j-\beta_k)\, v_{jkl}(\lambda) = 0, \quad k = 1,2,\ldots, b.$$

The rank of this system of equations is a+b-1. The first of these equations suggests the constraints

$$\sum_j \sum_k \sum_\ell \alpha_j \nabla_{jk\ell}(\lambda) = \sum_j \sum_k \sum_\ell \beta_k \nabla_{jk\ell}(\lambda) = 0 \tag{9.6}$$

be appended to produce a full rank system. Along with (9.6) we obtain the implicit equations

$$\mu = \frac{\sum_j \sum_k \sum_\ell x_{jk\ell} \nabla_{jk\ell}(\lambda)}{\sum_j \sum_k \sum_\ell \nabla_{jk\ell}(\lambda)} , \tag{9.7}$$

$$\alpha_j = \frac{\sum_k \sum_\ell (x_{jk\ell} - \mu - \beta_k) \nabla_{jk\ell}(\lambda)}{\sum_k \sum_\ell \nabla_{jk\ell}(\lambda)} , \quad j = 1,2,\ldots, a-1, \tag{9.8}$$

$$\beta_k = \frac{\sum_j \sum_\ell (x_{jk\ell} - \mu - \alpha_j) \nabla_{jk\ell}(\lambda)}{\sum_j \sum_\ell \nabla_{jk\ell}(\lambda)} , \quad k = 1,2,\ldots, b-1, \tag{9.9}$$

and

$$D = (1-2\lambda)^{-1} \frac{\sum_j \sum_k \sum_\ell (x_{jk\ell} - \mu - \alpha_j - \beta_k)(x_{jk\ell} - \mu - \alpha_j - \beta_k)^T \nabla_{jk\ell}(\lambda)}{\sum_j \sum_k \sum_\ell \left[\nabla_{jk\ell}(\lambda) - \left[\frac{1+2\lambda}{2+2\lambda}\right]^{\frac{1}{2}(p+2)}\right]} , \tag{9.10}$$

where

$$\nabla_{jk\ell}(\lambda) = \exp(-\tfrac{1}{2}(1+2\lambda)^{-1} (x_{jk\ell} - \mu - \alpha_j - \beta_k)^T D^{-1} (x_{jk\ell} - \mu - \alpha_j - \beta_k)). \tag{9.11}$$

Observations $x_{jk\ell}$ which require special consideration are indicated as in section 8, by low values of $\nabla_{jk\ell}(\lambda)$ vis-a-vis the whole set. A low value of $\nabla_{jk\ell}(\lambda)$ may mean that the particular observation is a potential outlier. Too many low values will imply

that the model assumption of a single Gaussian parent may not be warranted or that the model is mis-specified or that there are indeed a number of potential outliers. Furthermore, if $n_{jk} > 1$ and we find that individual cells have low values $v_{jkg}(\lambda)$ associated with them, then interaction in the table is a distinct possibility. In this case we generalize the model to

$$E(x_{jkg}) = \mu + \alpha_j + \beta_k + \gamma_{jk} .$$

and proceed accordingly.

This multivariate procedure can be especially useful for exploratory purposes. Determination of the sensitivity of $v_{jkg}(\lambda)$ and the parameter estimates to changes in λ will serve to uncover potential problems with the data or the working model considered as a consistent single entity. The procedure is computationally inexpensive and easy to use. We have investigated this procedure with respect to tests of hypotheses concerning the location parameters such as α_j and β_k but the distribution theory seems to be intractable.

Example 9.1. The data for this example concerning two-way cross classifications were taken from Anderson (1958, p. 218) who gives some additional background concerning these data. The first component of the observation vector is barley yield on a given plot in a given year; the second component is barley yield on the same plot made the following year. The treatments are five varieties of barley and we fit the model (9.1) to this data by the method of maximum likelihood and by the modified integrated squared error method for various values of λ with the objective of performing a sensitivity analysis. The results of this analysis are summarized in Tables 7 and 8. We have only given the results for λ=2 since the response of the parameter estimates and the final weights to decreases in λ continues the trend evidenced in Tables 7 and 8. Table 7 indicates that the largest change occurred in

the parameter α_s and the covariance matrix. The correlation increased from .22 to .37. The final weights $\nabla_{jk}(2)$ are given in Table 8. Observation (5,3) receives an especially low weight while observations (1,3), (3,4), (5,4), and, to a lesser extent, (2,4) also receive low weights. It is likely that the components of (5,3) are interchanged and should read (97,69) instead of (69,97). These observations have had the effect of reducing the correlation between the first and second year yields. We are suggesting that low weights raise the suspicion of potential outliers or other difficulties with the data and the model (Gaussianity, additivity, etc.), we are not suggesting, however, that this procedure be used as a formal test for outliers. Incidentally, tests of fit can often be used as or in lieu of tests for outliers.

Additional reduction of λ say to 1.5 will produce a somewhat stronger version of basically the same results. However, when λ is decreased to unity the procedure starts to deteriorate in the sense that the singled-out observations above receive weights near zero and a few of the other originally low weights have been further reduced. The first component variance is dramatically reduced. The reason for this is that the modified squared error multivariate residuals are becoming increasingly separated so that the empirical density estimator of the lack of fit distribution perceives multiple distributions and some of these are quite separated from each other.

A basic advantage of the modified squared error analysis of variance procedures is the direct adaptive involvement of the covariance structure in the development of estimates of model effects; this direct involvement of covariance structure considerably enhances our ability to constructively criticize (tentative) analysis of variance and regression models. We thus view the procedures and methods presented here as complementary and subordinate to, not as replacements for, the methods of least squares and maximum likelihood.

Table 7

Maximum Likelihood (ML) and Modified Integrated Squared Error (λ=2) Parameter Estimates

	μ	β_1	β_2	β_3	β_4	β_5
ML	109.1	-6.4	- 7.2	-5.6	18.4	0.8
	93.2	-7.0	-12.8	1.7	16.0	2.2
λ=2	108.8	-5.7	- 7.2	- 3.5	18.6	1.0
	92.8	-6.1	-11.2	- 1.6	17.9	4.2

	α_1	α_2	α_3	α_4	α_5	α_6
ML	- 6.3	46.5	-17.5	-17.1	-19.1	-20.9
	-10.4	23.6	25.8	- 1.4	-22.2	-15.6
λ=2	- 7.9	45.3	-19.7	16.9	-12.7	-21.4
	-10.8	22.3	25.3	- .1	-25.5	-16.1

$\hat{D}$

109.3	.22
26.7	133.9

$\tilde{D}(2)$

101.9	.37
42.4	125.5

Table 8

Barley Yield in A Given Year (1st tabular component on left), and Yield in Following Year (2nd tabular component on left), and Final Weights $\nabla_{jk}(2)(\times 100)$ (tabular component on right)

		VARIETIES									
		1		2		3		4		5	
Location	1	81 81	73	105 82	85	120 80	56	110 87	85	98 84	99
	2	147 100	94	142 116	82	151 112	100	192 148	66	146 108	88
	3	82 103	94	77 105	98	78 117	93	131 140	57	90 130	95
	4	120 99	87	121 62	68	124 96	98	141 126	77	125 76	69
	5	99 66	93	89 50	97	69 97	12	89 62	46	104 80	93
	6	87 68	95	77 67	98	79 67	94	102 92	98	96 94	87

Acknoledgement

The work of A.S. Paulson was supported by U.S. Army Research Office Contract DAAG29-81-K-0110.

References

1. Anderson, T.W. (1958). An Introduction to Multivariate Statistical Analysis. New York: Wiley.

2. Andrews, D.F. and Pregibon, D. (1978). Finding the outliers that matter. Journal of the Royal Statistical Society, B, 40, pp. 85-93.

3. Barnett, V. and Lewis, T. (1978). Outliers in Statistical Data. New York: Wiley.

4. Belsley, D.A., Kuh, E., and Welsch, R.E. (1980). Regression Diagnostics: Identifying Influential Data and Sources of Collinearity. New York: Wiley.

5. Box, G.E. P., (1979). Some problems of statistics and everyday life. Journal of the American Statistical Association, 74, pp. 1-4.

6. Bryant, J., and Paulson, A.S. (1979). Some comments on characteristic function-based estimators. Sankhya, A, pp. 109-116.

7. Campbell, N., (1980). Robust procedures in multivariate analysis I: Robust covariance estimation. Applied Statistics, 29, pp. 231-237.

8. Cramer, H. (1946). Mathematical Methods of Statistics. Princeton: Princeton University Press.

9. Daniel, C. (1978). Patterns in residuals in a two-way layout. Technometrics, 20, pp. 385-396.

10. Devlin, S.J., Gnanadesikan, R., and Kettenring, J.R. (1976). Robust estimation and outlier detection with correlation coefficients. Biometrika, 62, pp. 531-545.

11. Devlin, S.J., Gnanadesikan, R., and Kettenring, J.R. (1981). Robust estimation of dispersion matrices and principal components. Journal of the American Statistical Association, 76, pp. 354-362.

12. Dwyer, P.S. (1967). Some applications of matrix derivatives in multivariate analysis. Journal of the American Statistical Association, 62, pp. 607-625.

13. Gnanadesikan, R. and Kettenring, J.R. (1972). Robust estimates, residuals, and outlier detection with multiresponse data. Biometric, 28, pp. 81-124.

14. Gnanadesikan, R. (1977). Methods for Statistical Data Analysis of Multivariate Observations. New York: Wiley.

15. Huber, P.J. (1981). Robust Statistics. New York: Wiley.

16. Maronna, R.A. (1976). Robust M-estimators of multivariate location and scatter. Annals of Statistics, 4, pp. 51-67.

17. Parr, W.C. and Schucany, W.R. (1980). Minimum distance and robust estimation. Journal of the American Statistical Association, 75, pp. 616-624.

18. Parzen, E. (1962). On estimation of a probability density function and mode. Annals of Mathematical Statistics, 33, pp. 1065-1076.

19. Paulson, A.S. (1986). Generalized likelihood and model-critical procedures for data analysis and modeling. Submitted for publication.

20. Paulson, A.S., Roohan, P.J., Hwang, N.L., and Fuller, M. (1986). A test of goodness of fit for multivariate normality. Submitted for publication.

21. Paulson, A.S. and Swope, J. (1986). Information divergence, generalized likelihood, and tests of the composite hypothesis of p-variate normality. Submitted for publication.

22. Rosenblatt, M. (1956). Remarks on some non-parametric estimates of a density function. Annals of Mathematical Statistics, 27, pp. 832-837.

23. Watson, G.S. and Leadbetter, M.R. (1963). On the estimation of the probability density, I. Annals of Mathematical Statistics, 34, pp. 480-91.

MASTER COPY - FOR REPRODUCTION PURPOSES

UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER ARO-10	2. GOVT ACCESSION NO. N/A	3. RECIPIENT'S CATALOG NUMBER N/A
4. TITLE (and Subtitle) Some Integrated Squared Error Procedures for Multivariate Normal Data		5. TYPE OF REPORT & PERIOD COVERED Working Paper 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) A.S. Paulson C.E. Lawrence N.J. Delaney H.L. Hwang		8. CONTRACT OR GRANT NUMBER(s) DAAG29-81-K-0110
9. PERFORMING ORGANIZATION NAME AND ADDRESS Rensselaer Polytechnic Institute Troy, New York 12180		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS U. S. Army Research Office Post Office Box 12211 Research Triangle Park, NC 27709		12. REPORT DATE 13. NUMBER OF PAGES
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) Unclassified 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

NA

18. SUPPLEMENTARY NOTES

The view, opinions, and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy, or decision, unless so designated by other documentation.

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

multivariate Gaussian, parametric density estimation, tests of fit, outliers, influence functions, experimental design, cluster analysis, robustness

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Two methods of estimation for the parameters of the multivariate normal distribution based on the sample characteristic function are given. These methods are shown to have an equivalent basis in terms of Parzen kernel-like density estimation. The estimators for the mean vector and covariance matrix are dependent on a user-specified parameter. Variation of the user-specified parameter produces a response surface in the parameter estimates and therefore allows for an informal sensitivity analysis of the data with respect to a tentative working model. The informal

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

20. sensitivity analysis is intricately related to formal tests of fit of the model. The estimators of mean vector and covariance matrix have desirable robustness properties, are easy to compute and use, are relative efficient at the multivariate normal, and are useful in identifying potential outliers and inconsistencies in some statistical assumptions. These methods are directly applicable to structured data such as multivariate experimental designs. Several illustrations are provided.

END

5-87

DTIC